第 32 卷　第 20 期　　计　算　机　工　程　　2006 年 10 月
Vol.32　№ 20　　Computer Engineering　　October 2006
·安全技术·　　文章编号：1000—3428(2006)20—0170—03　　文献标识码：A　　中图分类号：TP393.08

基于验证码破解的 HTTP 攻击原理与防范

吉治钢
(网易互动娱乐有限公司，广州 510665)

摘　要：为防止基于表单自动提交的 HTTP 攻击，验证码技术得到了广泛应用。论文对常见的几种验证码形式作了简要介绍，讨论了验证码的破解原理，实验表明，互联网上相当多的验证码都不具有可靠的安全性。最后结合 OCR 技术探讨了一些防范方法。
关键词：验证码；HTTP 攻击；Internet 安全

Principles and Prevention of HTTP Attacks Based on Identifying Code Recogniztion

JI Zhigang
(Netease Interactive Entertainment Co., Ltd., Guangzhou 510665)

【Abstract】To avoid HTTP attacks using automatic form-committing, the identifying code technique is widely used. A brief introduction of the types of identifying code techniques and its application is given. The principles of recognizing and attacking are discussed. Primary experiments suggest that quite a lot of identifying codes are not secure enough. Finally, some methods and schedules with OCR techniques for prevention are proposed.
【Key words】Identifying code; HTTP attacks; Internet security

在网络飞速发展的今天，网速已不再成为网络访问的瓶颈，在为人们上网提供更快的访问速度的同时也给黑客们提供了更广阔的发展空间，在线破解对网络安全的威胁越来越大[1]。为了确保用户提交的请求是在线进行的正常操作，越来越多的网站都采用了验证码技术，以保证服务器系统的稳定和用户信息的安全。

当前，验证码的形式和样式都日益丰富，但其自身的安全性却为很多站点所忽略。本文讨论验证码的多种形式及其安全性问题，并用实验重点分析了图片验证码的破解原理，结果表明：相当多的验证码都面临着严峻的安全形势。

1 验证码

1.1 基于表单自动提交的 HTTP 攻击

根据 HTTP 协议，攻击者可以编写程序模拟表单提交的方式，将非正常的数据向网站服务器自动、快速提交，这就构成了基本的 HTTP 攻击。如图 1 所示，其中，虚线表示攻击者的数据自动提交方式。

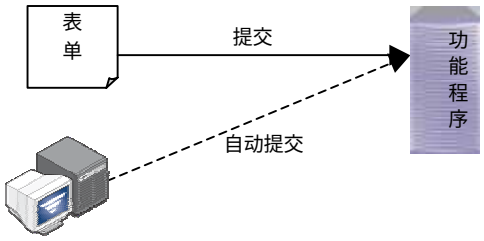

图 1　基于表单提交的 HTTP 攻击模式

由于互联网上涉及到用户交互的很多操作(如注册、登录、论坛发帖)都是用表单提交的方式实现的，因此，这种简单的 HTTP 攻击可能会导致以下几种安全问题：

(1)攻击者可以在短时间内注册大量的 Web 服务账户。这不但会占用大量的服务器及数据库资源，攻击者还可能使用这些账户为其他用户制造麻烦，如发送垃圾邮件或通过同时登录多个账户来延缓服务速度等；

(2)攻击者可以通过反复登录来暴力破解用户密码，导致用户隐私信息的泄漏；

(3)攻击者可以在论坛中迅速发表成千上万的垃圾帖子，严重影响系统性能，甚至导致服务器崩溃；

(4)攻击者可对系统实施 SQL 注射或其它脚本攻击，从而窃取管理员密码，查看、修改服务器本地文件，对系统安全造成极大威胁。

1.2 基于验证码的表单提交流程

为了防止攻击者利用程序自动注册、登录、发帖，验证码技术日益得到广泛的应用。基于验证码的表单提交流程如图 2 所示。

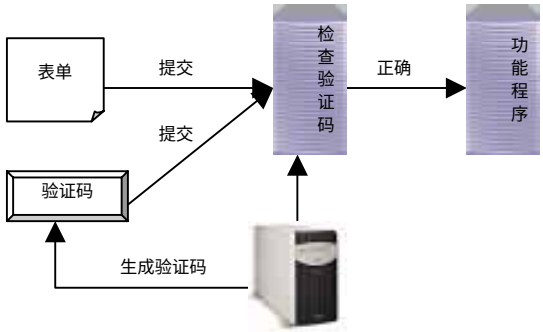

图 2　基于验证码的表单提交流程

作者简介：吉治钢(1978－)，男，硕士，主研方向：图像处理，网络与数据安全，入侵检测
收稿日期：2005-11-02　**E-mail：**syfool@188.com

所谓验证码，就是一串随机产生的字符串。服务器端将随机产生的验证码写到内存中，同时以某种形式展现给用户，用户在提交表单时必须同时填写验证码，如果与服务器端保存的字符串相同(即验证成功)，才能继续操作；否则，用户将无法使用后续的功能[2]。

对比图 1，可以看出，这种流程多了图片验证码的生成与验证机制。因此，验证码又称“附加码”。由于验证码是随机产生的字符串，每次请求都会发生变化，攻击者难于猜测其具体内容且无法穷举，模拟表单提交时便很难正确填写并通过验证，这样就实现了阻挡攻击的目的。

1.3 验证码的有效性

作者注意到，图 2 所示流程的有效性基于以下两个很重要的假设：

假设 1 用户可以收到并了解验证码；

假设 2 攻击者的自动程序无法了解验证码。

这二者必须同时成立。因为，如果用户不能了解验证码，那么将无法完成提交动作；如果可以编写程序自动获取验证码，那么攻击者就能够通过验证过程，实现攻击行为。

认识到验证码的有效性假设，对于验证码技术的评价、破解及防范技术都具有极为重要的意义。

1.4 进制验证码的类型

当前，互联网上较为常见的验证码主要有以下几种：

(1)文本验证码：在网页上以文本形式呈现给用户；

(2)手机验证码：用户在网页上提交自己的手机号码，系统以短信形式将验证码发送到用户手机上；

(3)邮件验证码：用户在网页上提交自己的电子邮箱，系统以 e-mail 形式将验证码发送到用户的邮箱中；

(4)图片验证码：又称“验证水印”，在网页上以图片形式呈现给用户。

2 对验证码的破解与攻击

尽管验证码对表单提交流程的安全起到了很重要的作用，但其自身的安全性却为很多站点所忽略，以致成为新的安全隐患。

2.1 文本验证码

由于验证码内容会原原本本地写在用户浏览到的网页中，编写程序对 HTML 文件进行一定分析后，同样可以获知验证码内容。因此，文本验证码的安全性很差，目前已经很少有网站采用这种形式。

2.2 手机验证码

由于需要查看手机才能知道验证码内容，攻击者通常没有办法实现自动获取，因此，仅从验证码的角度来说，这种方法可以较好地阻挡攻击者。

手机验证码的问题主要存在两点：

(1)受移动运营商短信网关的限制，有时会导致用户无法收到短信，从而使假设 1 不成立；

(2)可能造成对手机的 DoS 攻击：将指定手机号用于接收验证码，编写程序不断向服务器提交请求，就会使该手机不断收到验证码短信，对用户造成骚扰，甚至导致手机死机等后果。

因此，采用手机验证码时，通常需要与其它验证、防范手段相结合，才不致造成严重后果。

2.3 邮件验证码

这种形式的验证码仅仅比文本验证码的安全性略高，但仍然不能保证基本的安全性。原因有两点：

(1)利用 POP3 协议，可以编写程序从电子邮箱中获取电子邮件，经过对邮件的解码和文本分析，攻击者就可以自动取得验证码的内容了。

(2)与手机验证码相似，攻击者可以利用这种方式向被攻击者的电子邮箱发起 DoS 攻击，导致被攻击者的邮箱充满相关垃圾邮件，无法接收新邮件。

2.4 图片验证码

(1)对XBM格式的破解[3]

x-xbitmap 格式(以下简称为 XBM 格式)其实并不是真正的图片，而是一个纯文本文件，它需要由浏览器翻译并显示。XBM 文件只能显示黑/白两种颜色，而且要以数组的方式来表现每个要显示的图形，难以生成较为复杂的图案。

因此，攻击者可以很容易地根据文本信息了解到图片的构造，并以最简单、实用的验证码特征库(如数字的“腰粗”等)，快速识别出验证码的内容。

(2)基于模板匹配的破解

由于 XBM 图片验证码的安全性较差，很多站点都不再采用，而代之以真正的图像格式(如 BMP、JPG、PNG、GIF 等)。这样，验证码就是以点的方式而不是字符方式呈现给用户了，而且可以生成非常复杂的图案。因此，破解它需要完成以下两个关键步骤：1)解析不同的文件格式，得到所有像素点的信息；2)将点的信息转换为文本信息。

尽管各种图像文件格式都是公开的，但要从点阵中提取验证码信息却需要一些较为复杂的识别技术才能实现。

然而，当前很多图片验证码都具有一些对攻击者很有利的特点，如字符集元素较少(通常为数字或数字＋字母)；字体单一(通常为确定的一种印刷体)；字符大小通常大致相同。

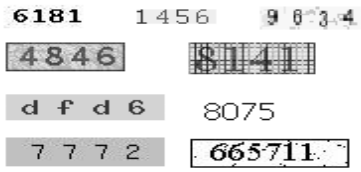

图 3 一些著名网站的图片验证码

如图 3 所示，这些验证码均取自国内著名站点(主要是一些著名的邮箱产品、论坛、以及门户)。容易发现，如果能够将图片中的各数字取出来作为模板，那么，通过与图片逐个比对就很有可能识别出字符串信息。为了验证这个想法，本文做了如下实验：

(1)使用图像处理软件手工将各数字提取出来，得到 10 个数字图片的模板库。

(2)要对图片验证码做预处理，这主要是为了消除背景混淆的影响。通过图像处理软件，可以发现，多数验证码的像素 RGB 值都与背景像素有很大差异，因此，在程序中做一些简单的判断即可将验证码从背景中分离出来。

实验中，本文将数字模板及预处理后的图片都保存为单色位图，这不但减小了模板匹配的算法复杂程度，也有效降低了识别的误差率。

(3)利用模板库进行验证码的识别。如果验证码各字符的坐标位置和字符大小都是固定的，那么，只需截取指定起点位置、指定长宽值的矩形区域与各模板逐个比对，匹配度最高的数字即为识别结果。

如果验证码各字符的位置不固定，那么可以采用滑动窗

口的办法：在指定区域内，滑动截取不同起点位置、指定长宽值的矩形，不断与模板相比对，匹配度最高的数字即为识别结果，对应的起点位置则为字符的坐标。

通过以上尝试，对图 3 中各验证码的识别结果见表 1(为安全起见，本文将网站名称隐去，仅以验证码作为统计标识)。

表 1　对部分图片验证码的识别结果

图片标识	验证码		识别率	
	图片个数	字符数	验证码	字符
6181	10	40	100%	100%
1456	10	40	100%	100%
9634	10	40	100%	100%
4846	10	40	100%	100%
8141	10	40	60%	85%
dfd6	10	40	100%	100%
8075	10	40	100%	100%
7772	10	40	100%	100%
665711	10	60	100%	100%

事实上，本文所用的方法只能识别同样大小、同种字体的字符，对于倾斜、笔画变粗(变细)等变化均无良好的适应能力[4,5]；然而，这种简单的方法却达到了较为理想的识别率。基于此，作者有理由认为：目前互联网上相当多的图片验证码都不具备可靠的安全性。

3 防范

对基于验证码的表单提交流程，主要有两方面需要注意：一是要避免直接利用验证码的进行攻击；二是确保验证码自身不易破解。

根据上述讨论和演示，本文认为：

(1)任何时候，都不应当使用安全性很差的文本验证码；

(2)应尽量不使用手机验证码、邮件验证码，以避免手机/邮件 DoS 攻击；

(3)建议使用安全性较高的图片验证码。

事实上，对图片验证码的识别与光学字符识别(optical character recognition，OCR)技术在本质上是完全相同的。而在OCR领域，目前对印刷体(数字、西文字母，甚至汉字)的识别技术已经相当成熟，对联机手写体的识别也已实现了商用[4]。因此，图片验证码面临的安全形势相当严峻。

考虑到目前OCR技术中尚存在一些不够成熟的领域，如脱机手写体的识别、多语言文字混排的识别、退化严重的文字识别等[5]，本文建议在图片验证码的设计中，加强以下几方面的变化：扩展字符集(可以考虑用汉字作为字符集)；随机变化字体和字符大小；随机设定字符的倾斜程度；随机设置字符坐标位置；增强背景混淆等。

当然，根据 1.3 节讨论的有效性假设，在增强图片验证码安全性的同时，还要注意不能使用户肉眼的分辨过于困难。

4 结论

为加强互联网应用的安全性，验证码技术得到了广泛应用，但验证码自身的安全性却为许多站点所忽略。

本文讨论验证码的各种形式及其安全性，重点介绍了图片验证码及破解原理，并用实验表明，相对多的验证码都不具备可靠的安全性。最后探讨了一些防范技术。但如何把图片验证码设计得既难于破解，又容易使肉眼分辨，尚需进一步的研究和探讨。

参考文献

1 红牌楼主. 论附加码在网络安全中的作用[EB/OL]. http://www.myrain.org/info/1234-1.htm, 2004-09-08.

2 陈十三哥. 解密验证码技术[EB/OL]. http://www.juntuan. net/jtyc/wenzhang/n/2005-07-30/6750.html, 2005-07-30.

3 Hemon. XMB 图片格式验证码的破解识别[EB/OL]. http://www.hemon.info/blogview.asp? logID=3, 2004-10-11.

4 陈友斌. 非特定人脱机手写汉字识别方法的研究[D]. 北京: 清华大学, 1996-07.

5 李　佐, 王姝华, 蔡士杰. 基于特征行必要-充分性匹配的字符识别方法[J]. 软件学报, 2002,13(1): 85-91.

(上接第 169 页)

$$(g^{x_1y_1})^{s_As_B} = g^{H(X)H(Y)}(Q_B)^{r_BH(X)}(Q_A)^{r_AH(Y)}(g^{ab})^{r_Ar_B} \bmod p$$

也无法计算出 K。因此，即使 A、B 的私钥泄露也不会对 K 构成威胁，所以协议具备前向保密性。

3.3 协议计算量分析

一个协议除了在安全上要满足一定条件，其效率也是一个需要考虑的因素，因此有必要对其计算量作一定分析。

本协议主要以 DSS 和 DH 协议为基础，这两个协议在计算上都以模幂运算为主，模幂运算非常消耗时间和资源，其比较好的解决方法是通过专用的密码处理器来计算。现在在硬件设计上一般采用 Montgomery 算法，这是一种比较好的解决模幂运算的方法。本协议与文献[3]提出的协议相比，A、B 双方虽然各多了一个随机参数，但因为双方只要计算一个密钥，所以模幂运算的总的次数并没有增加。

另外，从协议执行的步骤来看，要求协议的步骤和传递的参数尽可能少，这样有利于提高协议的效率和安全性。本协议与文献[3]中的协议都是 4 轮，传递的参数虽然因双方各多了一个随机参数略有增加，但总体的效率并没有太大改变。

4 结束语

本文介绍了一种将 DSS 签字标准与 DH 协议相结合的密钥交换协议，并指出其弱点，在此基础上提出了一种新的密钥交换协议并对其安全性和计算量作简要分析，可以看出新的协议能较好地抵抗已知密钥攻击、未知密钥共享攻击，并具有较好的前向保密性，而且在计算量上也没有太大改变。

参考文献

1 Arazi A. Integrating a Key Cryptosystem into the Digital Signature Standard[J]. Electronics Letters, 1993, 29(11): 966-967.

2 Nyberg K, Rueppel R A. Weaknesses in Some Recent Key Agreement Protocols[J]. Electronics Letters, 1994, 30(1): 26-27.

3 Harn L, Mehta M, Hsin W. Integrating Diffie-Hellman Key Exchange into the Digital Signature Algorithm(DSA)[J]. IEEE Communication Letters, 2004, 8(3): 198-200.

4 ElGamal T. A Public Key Cryptosystem and A Signature Scheme Based on Discrete Logarithms[J]. IEEE Trans. on Information Theory, 1985, 31(14): 469-472.

5 Kaliski B. An Unknown Key-share Attack on the MQV Key Agreement Protocol[J]. ACM Transaction on Information and System Security, 2001, 4(3): 275-288.

6 王育民, 刘建伟. 通信网的安全——理论与技术[M]. 西安: 西安电子科技大学出版社, 1999.